Recommendation techniques in forensic data analysis: a new approach

M. Quintana, S. Uribe, F. Sánchez, F. Álvarez

Keywords: Forensic analysis, recommendation, collaborative filtering, suspects, clues.

Abstract

Data mining for digital forensic analysis is a branch of Computer Science focused on pattern extraction from large-scale data which has been used to support analysts when trying to solve crimes. One of the most promising applications of data mining algorithms is to build recommendation systems, aiming to propose future directions to the investigation and to guide the analyst through the process.
In this paper we propose a new approach, architecture and framework with the purpose of taking advantage of the recommender systems techniques to the forensic field and provide examples of their applicability to different use cases involving large scale collections of multimedia information related to a defined forensic case.

1 Introduction

Police investigations when involving large scale collections of data are assisted by data mining systems for a digital forensic analysis. This branch of Computer Science is focused on patterns extraction from large-scale data which has been used to help analysts when trying to solve crimes. Applications of this kind of algorithms are focused on making simpler investigations for experts on criminology, such as recommendation systems for forensic analysis. The aim of this paper is to propose an approach to take advantage of the recommender systems techniques for digital forensic analysis. In this paper we present a new framework based on recommendation techniques using large multimedia data collections (text, audio, images, video), to increase efficiency and effectiveness of the forensic analysis, compared to other similar approaches.

Recommendation systems are widely used in our society mainly for media content applications (such as recommending pieces of content to users from a large content collection), and considering a similar kind of content used in digital forensics, we have found some analogies which can be exploitable. Recommendation systems are classified depending on the features they rely on: content-based, collaborative filtering based and hybrid [1]. The first kind rely on the attributes (technical, semantic…) of the available content, the second one on previous users satisfaction, being the latter a mix them. In our work we rely on previous cases knowledge to correlate clues and context from different investigations.

Correlation of hints and context information from the scenario of the crime are the conceptual basis to infer proper knowledge on specific cases. Tool development will provide support to the police officers/analysts in order to indicate the most likely directions of the investigation.

The rest of the paper is organized as follows: Section 2 presents the current state of the art of previous published systems. In section 3 the components of the system and how they work together are detailed. Different scenarios where the system could have an application are described in section 4. Finally, the conclusions and future research lines extracted from this work are discussed in section 5.

2 Related work

The application of data mining techniques in forensic investigations represents an important field of study. Accordingly, over the last few years, computational processes have been applied in order to improve police investigations in different aspects, such as the development of efficient tools for clue analysis, the management of vast amounts of data, and the implementation of artificial intelligence abilities for reasoning [2].

In this framework, clustering algorithms are one of the most useful techniques for crime data analysis. Different researches such as [3] and [4] are in charge of proving effectiveness on crime patterns definition and identification, which is especially relevant in forensics research. Moreover, other techniques such as fuzzy methods are applied as well to develop expert systems for forensic data analysis [5]. Several solutions proposed in a general way [6] [7] by the fuzzy research community should be taken into account to deal with the uncertainty and volatility of the scenario.

Additionally, outside the forensic field, social recommenders are mainly based on the current technique of memory-based collaborative filtering. That technique covers two analogous solutions, user-centered and item-centered approaches, both recommending items to users. The difference is that the first solution applies similarity metrics among users and the second applies similarity metrics among items. The item-based collaborative filtering [8] uses a rating matrix to compute pairwise similarities among the items, and it stores the top 20 to 50 most similar items per each one in the similarity matrix. Finally, to create the prediction, the algorithm uses a weighted sum over all items similar to the unknown item that has been rated by the current user.
There are two problems related to item-based collaborative filtering:

- The assumption that a rating is defined by ratings for commonly co-rated items from all the users is hard to justify.
- A lack of bias correction, every co-rated item is isolated.

Some new approaches are exposed in [9] like Model-based Collaborative Filtering. This technique was widely exploited at the Netflix prize [10], and it is based on the idea that the ratings are deeply influenced by a set of factors that are very specific to the domain. These factors do not use to be obvious, and the goal is to infer those so called latent factors from the rating data by using mathematical techniques. Other detected problems were related to the computation of large scale data [11] or the way to deal with the implicit feedback data [12].
Currently, memory-based techniques are still very popular, especially in some commercial systems, as said by Tang et al. [13]. However, there are a huge set of model-based algorithms very useful in real systems, thanks to the new techniques in parallelization, cloud computing and big data frameworks. Those solutions become important in scenarios where the algorithm has to deal with a very specific domain or with a set of users with enclosed features. Some of the most representative model-based collaborative filtering (CF) techniques are Bayesian Belief Nets CF, Clustering CF, MDP-based CF, Latent semantic CF, Sparse Factor Analysis, and CF using dimensionality reduction methods, such as SVD or PCA [13]. The most important issue in the design of a recommender system is to understand the nature of the data and decide what kind of technique is more suitable. In a second step, specific algorithms encompassing the presented techniques should be selected.

3 System Architecture

The proposed architecture presents two main functionalities, which can be used together or individually. As we can find in Fig. 1, these two functionalities are the recommendation of clues (to suggest the investigator the best next steps to solve the case) and the recommendation of suspects. Both functionalities are based on the information provided by past cases, and they are built over recommendation systems algorithms. Specifically, model based collaborative filtering methods inspire the "Clues Recommendation Algorithm" while memory based collaborative filtering techniques are the base of the "Suspects Recommendation Algorithm".

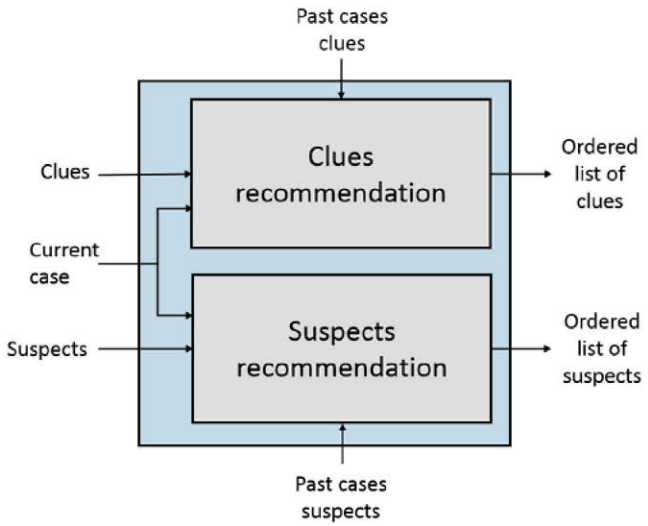

Fig. 1. Functionalities of the system

3.1 Clues Recommendation Algorithm

This functionality is especially relevant at first steps of the investigation, when there is a lot of information and the system can save some time, creating automatically the first suggestions. The complete flow of the algorithm is depicted in Fig. 2. The Clues Recommendation functionality starts from the information of the features that characterize a case. These features are modelled as clues that the investigator can follow to improve the investigation.

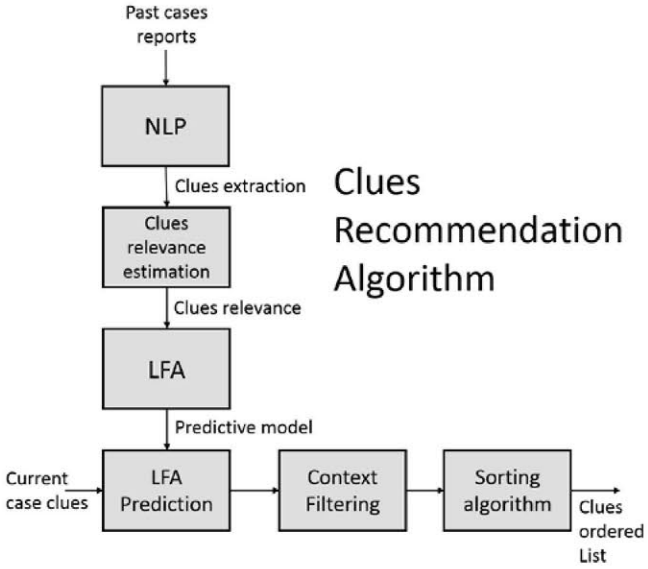

Fig. 2. Clues Recommendation Algorithm

Transformation from past cases information to relevant clues depends on the kind of source that is available. If we are using textual reports, a NLP (Natural Language Processing) module is needed to do the extraction. If we have images or

video files instead of reports, we would use an image or video analysis module replacing the NLP. The NLP functionalities that can provide the essential clues are based in well common techniques: named entity recognition [14] and subjects detection [15]. The application of both techniques to the body of the reports allows to extract the main clues in key-value pairs.

Relevance of clues in each case is estimated in the next module, the "Clues relevance estimation", using relevance algorithm such as tf-idf [16]. The information at this point is expressed as:

$$< c_m, l_n, r_{m \cdot n} > \quad (1)$$

Where c_m are the m past cases available, l_n are the n different clues extracted from the past cases, and $r_{m,n}$ expresses the relevance of each one of the n clues regarding to the n past cases. Therefore, the relevance of the clues in the cases is expressed using a $R_{m,n}$ matrix. This matrix is decomposed into the product of a case feature and a clue feature matrix. Each row in $C_{m,p}$ is the vector of affinity from every case to the features, while each row in $I_{p,n}$ is the vector expressing relation between clues and features. Decomposition is performed in the LFA (Latent Factor Analysis) module, where common techniques like Lanczos method for SVD are not applicable, because the relevance matrix is sparse and partially defined, and missing entries cannot be interpreted as 0.

$$\begin{pmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mn} \end{pmatrix} = \begin{pmatrix} c_{11} & \cdots & c_{1p} \\ \vdots & \ddots & \vdots \\ c_{m1} & \cdots & c_{mp} \end{pmatrix} \cdot \begin{pmatrix} l_{11} & \cdots & l_{1n} \\ \vdots & \ddots & \vdots \\ l_{p1} & \cdots & l_{pn} \end{pmatrix} \quad (2)$$

Decomposition must be performed using only known entries, and the objective is to find the set of cases and clues feature vectors that minimize the squared error to the known values:

$$\min_{U,M} \sum (r_{ij} - l_j c_i)^2 \quad (3)$$

The model contains hyperparameters, regularization and learning rate that need to be chosen. The regularization parameters are very important, because it is essential to avoid overfitting, which is a consubstantial problem of this kind of data.

The regularization is introduced using λ parameter and norm of the vector c_i and l_i. Besides, the t (i, j) parameter is included to adjust the confidence of each value, as it is shown in the following expression:

$$f(C, L) = \sum t(i,j)(r_{ij} - l_j c_i)^2 + \lambda(\sum \|c_i\|^2 + \sum \|l_i\|^2) \quad (4)$$

Chosen technique to solve modelling is the Alternating Least Squares, a widely used algorithm in the Latent Factor Analysis of collaborative filtering algorithms. This model is used in the "LFA Prediction" module to predict the relevance of the possible clues in the current case.

$$\widehat{r_{c,l}} = f(c, l) \quad (5)$$

Therefore, a list of relevant values for most probable clues is assigned to the current case. Next module "Context Filtering" filters the results removing clues and modifying the weights of its values depending on specific features such as location or time of the current case. It is performed using a rule-based algorithm. Finally, the "Sorting Algorithm" sorts the clues depending on its final relevance and provide the results to the investigator.

3.2 Suspects Recommendation Algorithm

Second functionality is based on a memory-based collaborative filtering algorithm, as in this case, latent factors could not be useful to predict possible suspects. The features of the suspects are very specific of each person, therefore the patterns could not be relevant, and a memory-based algorithm is more effective. The Algorithm is based on this idea: some of the possible suspects of the current case could be involved in the past in other cases similar to the current case. Therefore, a filter of the most probably known suspects is a very useful tool for the investigators, saving a lot of time by means of a probabilistic weighting.

Besides, when the investigator is dealing with a new crime situation, where no obvious suspects are likely to be found, the suggested suspect could be taken as clues to suggest probably features of the real involved suspects. The KNN intermediate stage is useful to this purpose, providing the weighted neighbourhood among cases.

The Fig. 3 shows the process of the "Suspects Recommendation Algorithm", which shares some common parts with the Clues Recommendation Algorithm.

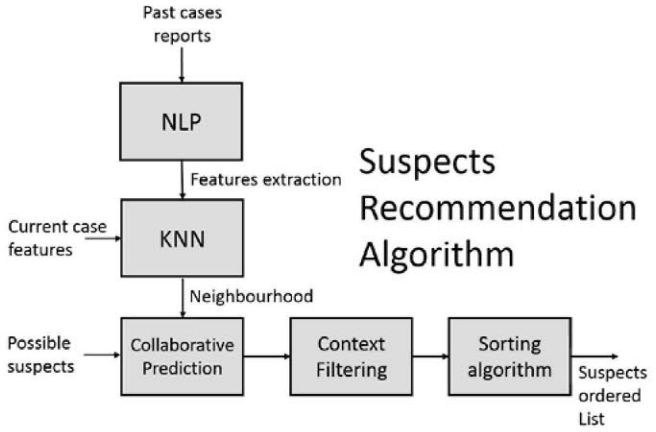

Fig. 3. Suspects Recommendation Algorithms

In this functionality, involved vectors are defined as:

$$< c_m, s_n, r_{m \cdot n} > \quad (6)$$

Where c_m are the m past cases available, s_n are the n different suspects extracted from the past cases, and $r_{m,n}$ expresses the involvement of each suspect in the n past cases.

After features extraction by using the NLP module (or analogue modules for image, video and audio), the most important step is the calculation of the neighbourhood of the current case. It is performed by means of a KNN (K-nearest

neighbours) algorithm, which calculates the K most similar cases to the current one using a model that has been previously learnt from similarities and the prediction success of other cases.
Using the neighbourhood of the current case, a common Collaborative prediction is performed, following this expression:

$$r_{c,s} = \bar{r_s} + \frac{\sum sim(c,c\prime)(r_{s,c\prime} - \bar{r_{c\prime}})^2}{\sum |sim(c,c\prime)|} \quad (7)$$

Where $r_{c,s}$ is the probability prediction of connection between the possible suspect s and the current case c, $\bar{r_s}$ expresses the mean of the involvement of the suspects s in past cases, sim(c,c') expresses the similarity between the current case c and the past case c', $r_{s,c'}$ expresses the involvement of suspect s in the past case c', and $\bar{r_{c\prime}}$ is the mean of the involvement of suspects in the case c' (the higher is this value, the more suspects have been involved in the case and, consequently, the relevance of each one should be less important in a relative way).
Finally, "Context Filtering" and "Sorting algorithm" modules are applied to obtain final suspects list, in an analogous way to the first functionality.

4 User cases application

Three challenging Use Cases (UCs) have been designed by expert end-users in LASIE project in order to cover the most important scenarios and to test all the implemented modules. In this respect, these UCs also represent three complete and advantageous environments to apply the recommendation system defined in this paper, due to their characteristics and objectives.

4.1 Use Case definition

In particular, the UCs are defined as follows

- Use Case 1.

This UC is focused on the analysis of civil disorders known as riots, which involves vandalism and destruction of public and private property. In particular, this scenario shows massive events such as peaceful protests, meetings or sport events which suddenly turns into a riot, mainly due to agitators among the participants. In this case, main evidences come from videos captured at the site of the incident, including media content from security systems surveillance cameras and personal videos captured by particular citizen.
The main objective in this case is to help the police with identification of special roles such as agitator one, which is usually the leading person of the riot. This can be achieved by means of the combination of two main modules: the first one focused on video processing for extracting the media evidences and the second one in charge of inferring which the main suspects are. The recommendation system contributes in this second stage of the investigation, as it is will be explained in the next sections.

- Use Case 2.

The second UC is related to accidents at workplaces, particularly at construction sites, where the main objective is to investigate if it is a real accident or if it is criminal responsibility of people involved (the injured person, the responsible for safety in workplace, etc.).
In this case, the investigation is usually focused on the analysis of different aspects of the scene, such as the visual inspection of the area, the identification of involved people, etc. Moreover, these evidences are gathered into a textual report which includes all clues obtained by the investigators. By analyzing the information, a police technical report can be obtained to objectively determine criminal liability.

- Use Case 3.

Finally, this UC is focused on the investigation related to a missing person. It is based on the audio and textual analysis from web and media resources (mobile calls, text messages, social networks activity, internet activity, etc.) in order to obtain different clues which can help in the case resolution.

4.1 Recommendation in Use Cases

Once the UCs are defined, next step consists of explaining the application of the recommendation module as a new tool for enhancing the case resolution.

- Suspects recommendation module in Use Case 1

The main objective of this case is to identify important person roles in riots scenarios, specially the agitator or leader of the disorders. For this purpose, the investigators make use of different video evidences to detect possible suspects.
According to this, the application of suspects recommendation module is particularly suitable to improve the investigation process. Dataflow for this UC is organized in some different steps relying on the architecture shown in Fig. 3. First step is based on feature extraction from past cases reports, analyzing characteristics of current cases. Then, from this analysis the system obtain a set of k similar cases previously solved. This neighbourhood is applied to all possible suspects detected by the police for a collaborative prediction. After that, obtained results are filtered using context features of the specific case in order to delimit possible suspects according to different parameter of the riot scenario such as place, date, event type, etc. Finally, a sorting algorithm is applied to obtain an ordered suspect list that can be used to facilitate and improve their investigations in terms of time consuming efficiency.
Although this scenario is specially focused on the suspects investigation, if there were other evidences or modules taking part in the investigation (such as an event detector module), it could be also applied the clue recommendation module to support the police work.

- Clue recommendation module in Use Case 2

The investigation in this UC is focused on the analysis of different textual evidences in order to determine the main

responsibilities of the accident. For this reason, application of clue recommendation module in Fig. 2 can relief investigators to detect possible skills to solve the case. In this respect, first step is formed by a detailed analysis of previous similar cases solved in order to obtain information about the main clues that were achieved at the investigation. Then, a LFA algorithm is applied, and a prediction model for new cases is obtained. This model is applied to current case clues, determining which are the most relevant according to the case characteristics and context. Finally, the system obtains an ordered list including all the clues based on its type and relevance.

- Clues and suspects recommendation module in Use Case 3.

Finally, and according to this UC definition, application of the entire architecture expressed in Fig. 1 can provide several advantages in relation to case resolution. In this way, clues recommendation module can be applied to determine which ones are the main textual and audio evidences according to the relevance of their NLP analysis (based on previous similar crimes). Then, analyzed evidences can provide a set of suspects, and the sorting algorithm provides an ordered list with the possible suspects according to the correlation of their features and case characteristics.

5 Conclusions and future work

Recommendation systems are suitable for heterogeneous data sources, as those collected for criminal investigations. We presented an architecture applying specified techniques with two main purposes: recommend the most relevant items among all collected information and suggest possible directions for the investigation based on the evidences and past cases. Context information will also be used to determine the boundaries of digital data analysis.

Massive performance results analysis will be the next step of the work that is being carried out. Data collection is on process to acquire a large data set covering all possible scenarios that will allow a proper evaluation of the developed system.

Once this step is completed, information managed by the system will be improved by implementing some high level knowledge inference algorithms that could rely on fields of science like Computer Vision or 3D scene analysis. Using these techniques, semantic concepts would be introduced on the data managed, and more accurate results would be obtained.

Pilot experience could finally provide a more reliable application on the field where this work is located. To define these events different investigators are being queried and its feedback is being used to properly deploy the system in a real scenario. Tests developed there should also include some confidence output from the system, which will make its results more trustable.

Acknowledgements

This publication is based on work performed in the LASIE project, funded by the Seventh Framework Programme of the European Union (607480 Grant Agreement).

References

1. J. Bobadilla, F. Ortega, A. Hernando, and A. GutiéRrez. 2013. Recommender systems survey. *Know.-Based Syst.* 46 (July 2013), 109-132.
2. Franke, K., Srihari, S. N.: "Computational Forensics: An overview". Computational Forensics. Lecture Notes in Computer Science. Volume 5158, 2008, pp1-10.
3. Nath, S. V.: "Crime pattern detection using data mining". In Proceedings of the IEEE/WIC/ACM International Conference on Web intelligence and intelligent Agent Technology, pp 41-44, 2006.
4. Bharathi, A. Shilpa, R.:"A survey on crime data analysis of data mining using clustering techniques". International Journal of Advance Research in Computer Science and Management Studies. Volume 2, Issue 8, 2014.
5. Stoffel, K., Cotofrei, P., and Han, D., "Fuzzy methods for forensic data analysis", IEEE International Conference of Soft Computing and Pattern Recognition, pp. 23-28, 2010.
6. Fu and Shen (2010) - Fuzzy Compositional Modeling, IEEE Transactions on Fuzzy Systems, vol. 18(4), pp. 823 - 840;
7. Lim and Chan (2015) - A weighted inference engine based on interval-valued fuzzy relational theory, Expert Systems with Applications, vol. 42(7), pp. 3410-3419.
8. Sarwar et al.: "Item-Based Collaborative Filtering Recommendation Algorithms", 2001.
9. Koren et al.: "Matrix Factorization Techniques for Recommender Systems", 2009
10. http://www.netflixprize.com (last accessed on 12/06/2015)
11. Zhou et al.: "Large-scale Parallel Collaborative Filtering for the Netflix Prize", 2008.
12. Hu et al.: "Collaborative Filtering for Implicit Feedback Datasets", 2008.
13. Tang, J.; Hu, X.; Liu, H., "Social recommendation: a review", *Social Network Analysis and Mining*, vol.3, no.4, pp.1113-1133, December 2013.
14. Rahul Sharnagat, "Named Entity Recognition: A Literature Survey", June 30, 2014, Center For Indian Language Technology.
15. Temizer, M.; Diri, B., "Automatic Subject-Object-Verb relation extraction," *Innovations in Intelligent Systems and Applications (INISTA), 2012 International Symposium on* , vol., no., pp.1,4, 2-4 July 2012
16. Wu, H. C., Luk, R. W. P., Wong, K. F., & Kwok, K. L. (2008). Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS)*, *26*(3), 13.